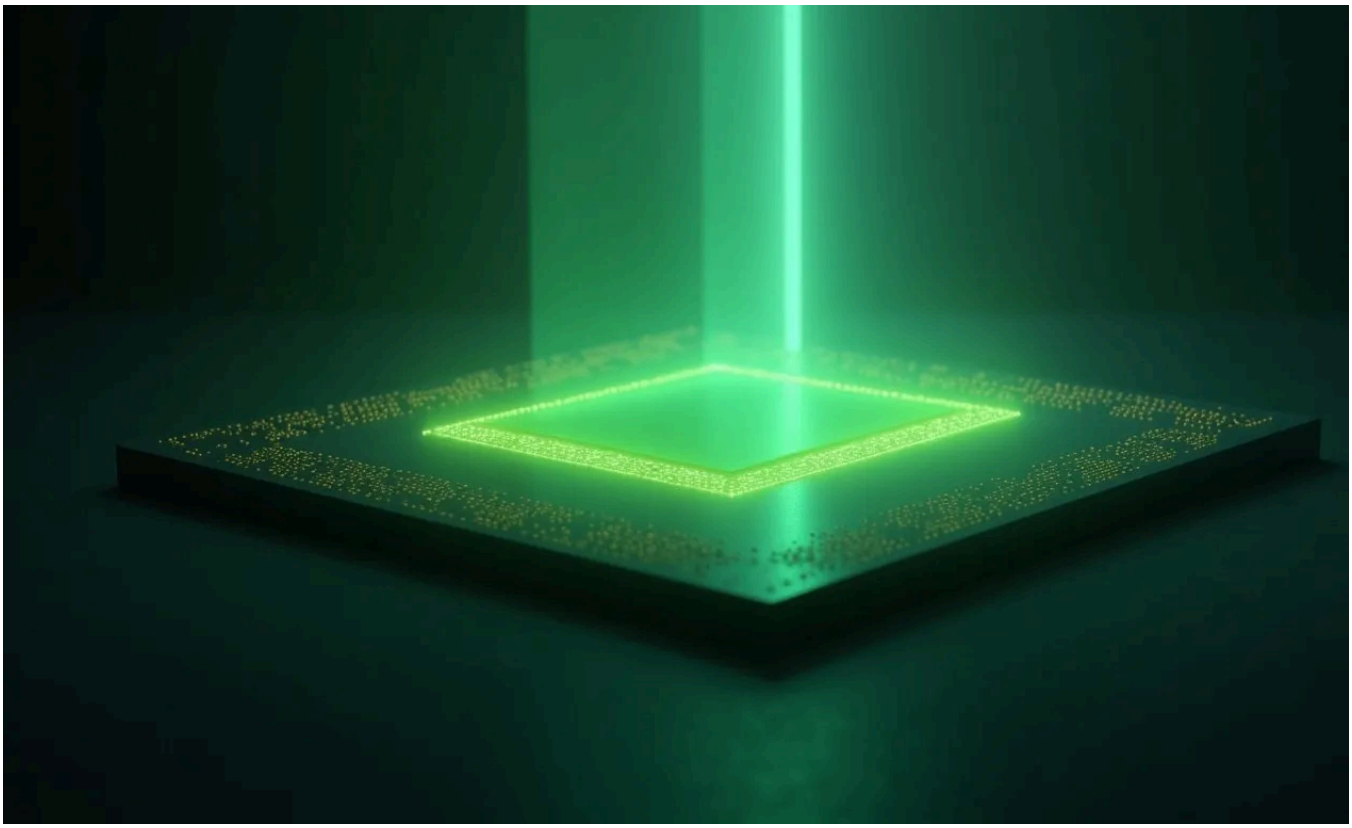
September 2024

Intel concedes training, Nvidia recovers from delays, and the race for HBM4 is on

HITESH KUMAR

SEP 30, 2024



Is this what we imagine on-chip free-space optics might look like?

This month's updates:

- Nvidia B100/200's not delayed? How TSMC and Nvidia overcame a major setback
- Intel concedes on training, focuses on AI inference for their next GPUs/X
- SambaNova using a single rack to achieve the fastest LLM inference spe to date
- Tesla develops its own transport protocol for its 100,000 GPU supercompu

- The race towards HBM4, a tie between Samsung and SK Hynix, Micron catching up

One-pagers:

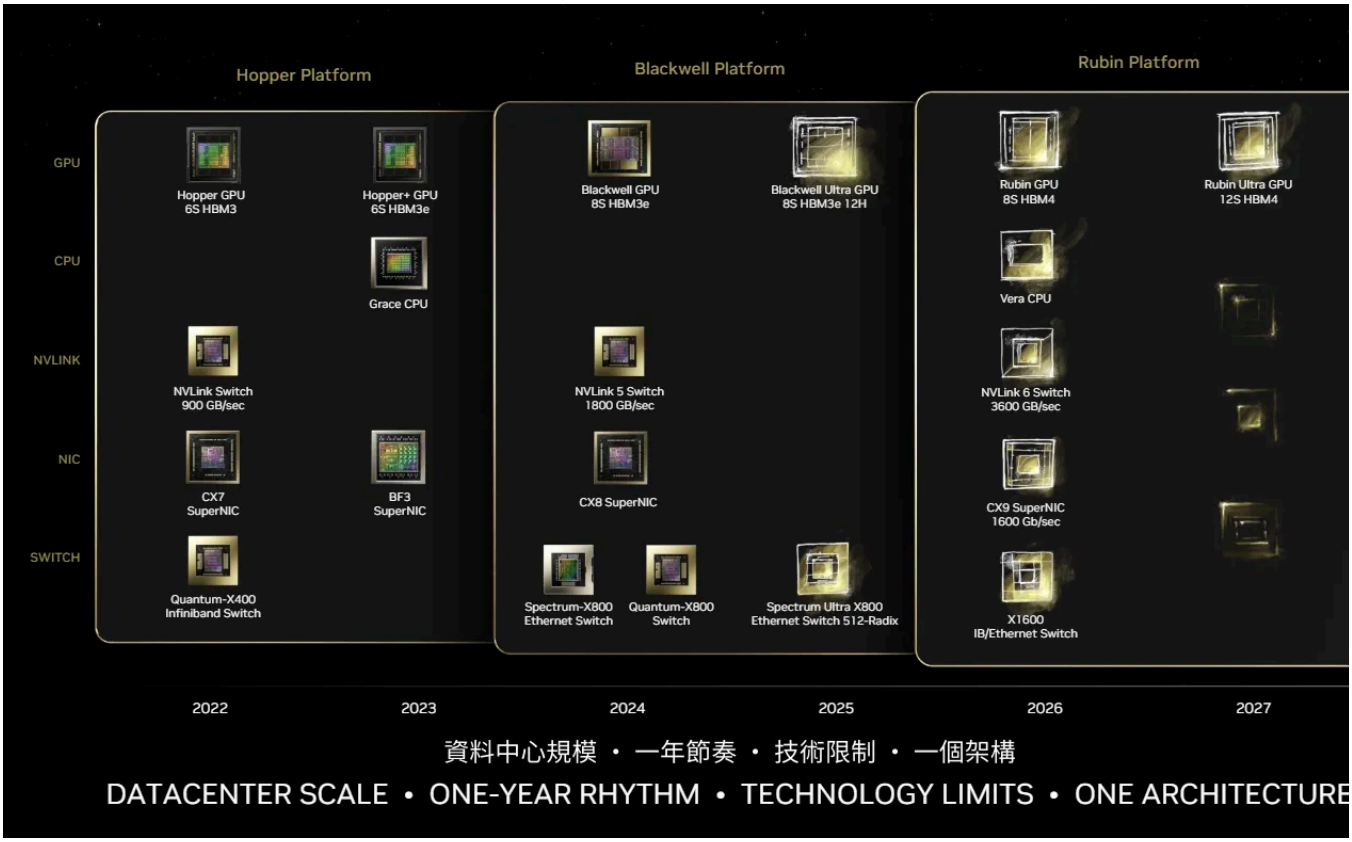
- LPDDR5X
- OSFP
- FLOPs

This month's updates:

Nvidia B100/200s not delayed?

Last time we reported that Nvidia's B100/200 cards would be delayed coming market, up until 2Q2025 at the latest from their original 4Q2024 planned shipments, but it appears that things are back on track, with TSMC and Nvid overcoming manufacturing issues and picking up pace to meet their contrac (almost)

The roadmap as shown below details their plans for their other technologies such as their switch ASICs. The Quantum and Spectrum series supporting up 144 and 64 ports at 800Gb/s respectively, with a 512-radix version of the spectrum coming in early 2025 hopefully, and the CX8 SuperNIC to support th bandwidths should, together, enable networks that will saturate the GPUs of t near future.



Source: Nvidia

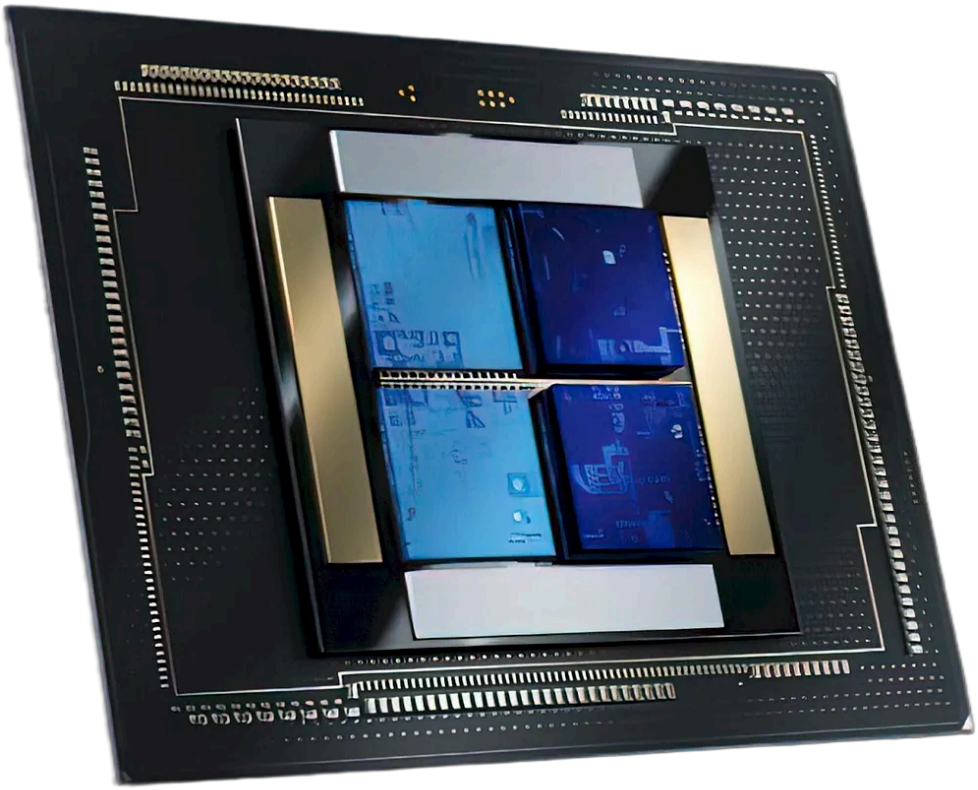
With individual orders being in the $100's of millions from some of their customers, Nvidia has had a lot of pressure to deliver their GPUs on time, but with the recent issues in TSMC's next-gen manufacturing process potentially delaying the production of these devices by months, things were looking blea

However, it appears that the issue has been solved, with Nvidia and TSMC making modifications and falling back to a more mature and tested semiconductor manufacturing technology to meet deadlines. The first produ out will be their racks, the NVL36 and NVL72, ultra dense NVLink connected setups that are designed to fit into modern datacenters with high power dens and liquid cooling. The B200A (lower power, lower cost version) and corresponding 36 and 72 GPU racks, all air-cooled, are still expected to ship a originally planned.

Intel concedes on AI training

A letter from Intel CEO, Pat Gelsinger, states that Intel's XPU lineup will be focusing on AI inference only, conceding AI training to its competitors. Whe

this is yet another failure on Intel's part to deliver, or a strategic move to focu their strengths, the Intel Falcon shores is still a capable AI accelerator.



Source: Intel - A TDP of 1500W, carrying 288GB of HBM3, 9.8TB/s of bandwidth between device memory and compute, with x86 (of course) cores on chip.

After the recent failures of the company, Intel has confirmed that its planned Falcon shores XPU (CPU + GPU on the same chip) will now be a GPU only an will focus on AI inference instead of training. Whilst it was known as far bac 2Q2023 that Falcon shores, Intel's own successor to the Habana Gaudi 3, will longer be a CPU and GPU on the same chip, only recently has it been made cl that Intel no longer plans to optimise its GPU and tech stack for AI training.

Not supporting AI training workloads will serve more than one purpose: designing the chip will be easier, the software stack will be easier to develop, the cost of production, including labour and manufacturing, will likely be significantly lower.

The TDP (thermal design power), or maximum amount of heat the device car output during typical workloads, is stated to be 1500W, 50% more than even tl Nvidia B200. Such a design will require a huge leap in socketing technology, exceeding the wattage that Nvidia's SXM socketing tech stack must support. addition, using these devices will force any datacentre to have state of the ar liquid cooling solutions, which as we have seen, is not an easy ask.

SambaNova's LLM inference speeds

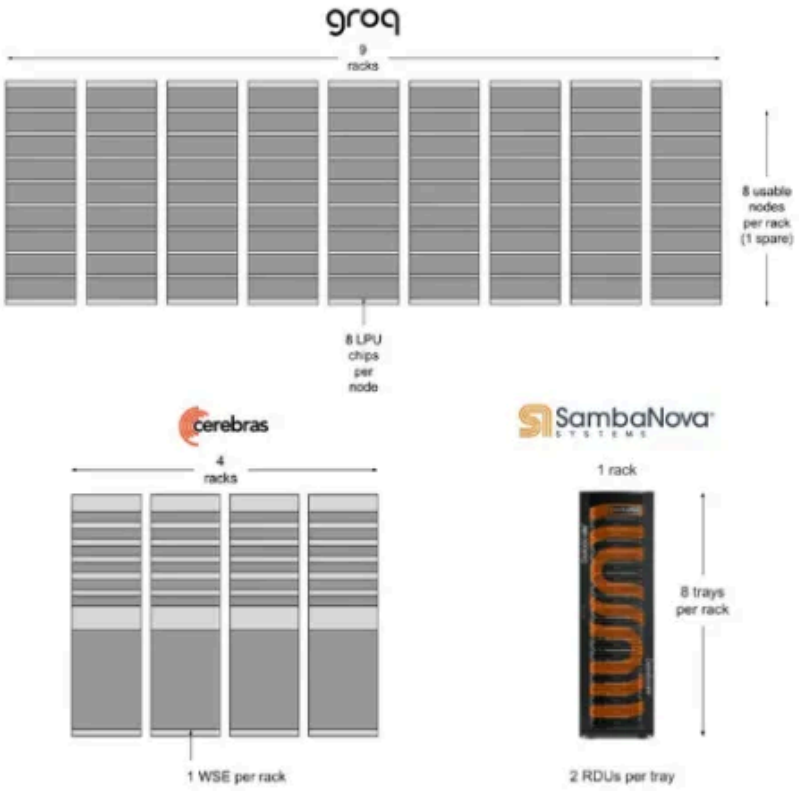
Despite providing little detail on the connectivity and power details of their A inference system, SambaNova systems and their SN40L AI accelerator are st very interesting to study. AI inferencing and breaking into the Middle east's growing AI hardware market are both tasks that SambaNova are setting recc in.

SambaNova systems, a SoftBank and BlackRock funded AI accelerator syste designer has been making headlines in the AI inference space with its recor breaking LLM inference speeds. Most companies who design and integrate systems that compete with the likes of Intel, Nvidia, AMD and others usually the open-source, open-weight (anyone can download and modify) Llama LL from Meta, since they can optimise and inference the models in any way the like to suit their own devices and setups. SambaNova has just reported speed up to 132 Tokens (for the purposes of this article, words)/second, for the 405 Billion parameter Llama 3.1 model.

Source: SambaNova

To put that into context, AWS and Azure both, likely serving the models on Nvidia H100s, offer speeds of 13 Tokens/second. What's even more surprising how space and energy efficient their system is, with the 16 x SN40L RDU (reconfigurable data unit, their AI accelerator) rack being sufficient to run a model this size.

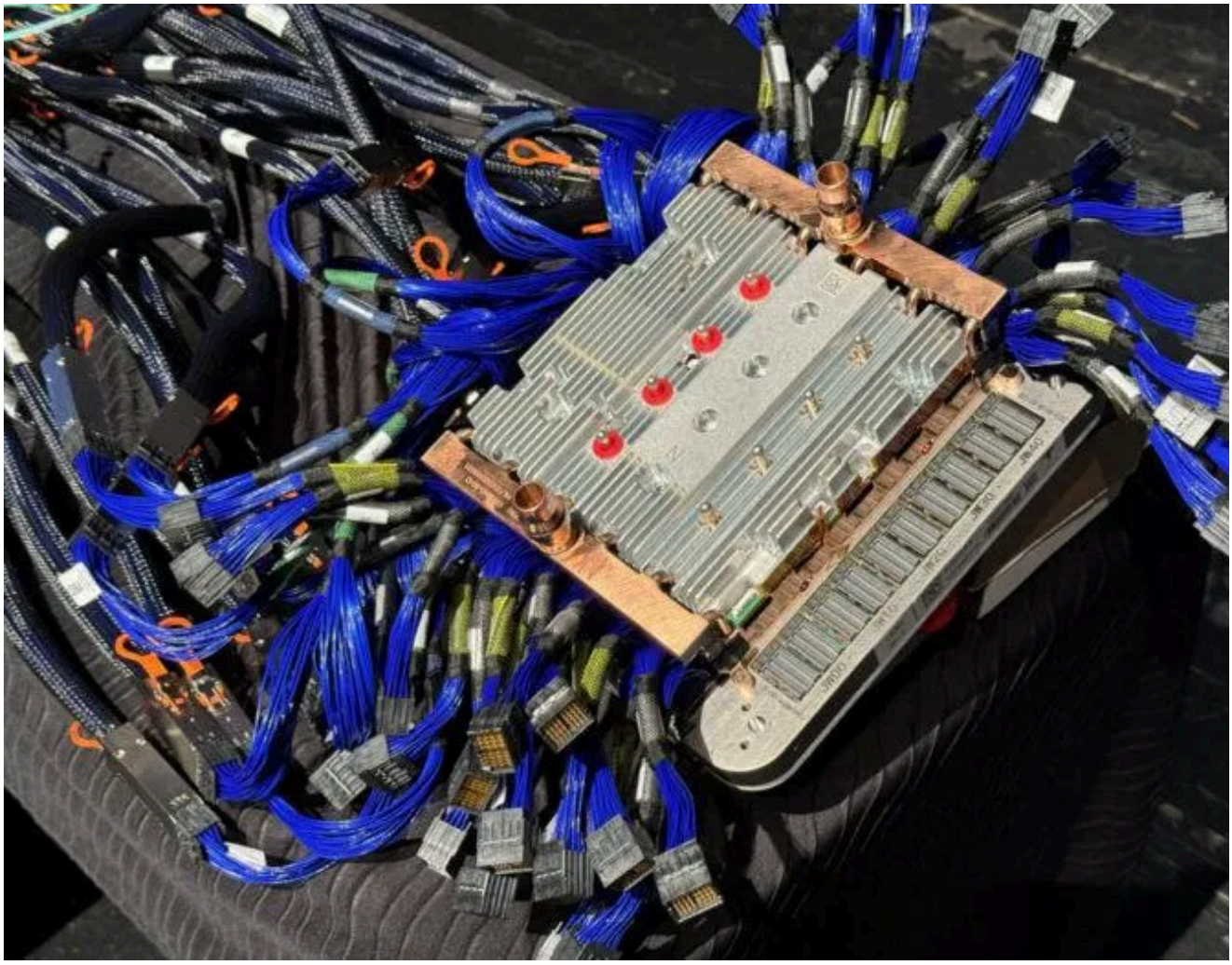
The RDU comes in at 51 Billion transistors per piece, compared to the H100s 8 Billion, and has 64 GB of HBM3, compared to the H100s 80/96 GB. Each of the trays in the rack support 2 RDUs with an additional 1.5 TB of DDR5 RAM (1). T real secret through, of device interconnectivity, whether it be via copper like Nvidia's NVLink or via optics like an Ethernet network, is kept from the publ



Source: SambaNova

Tesla's transport protocol over Ethernet

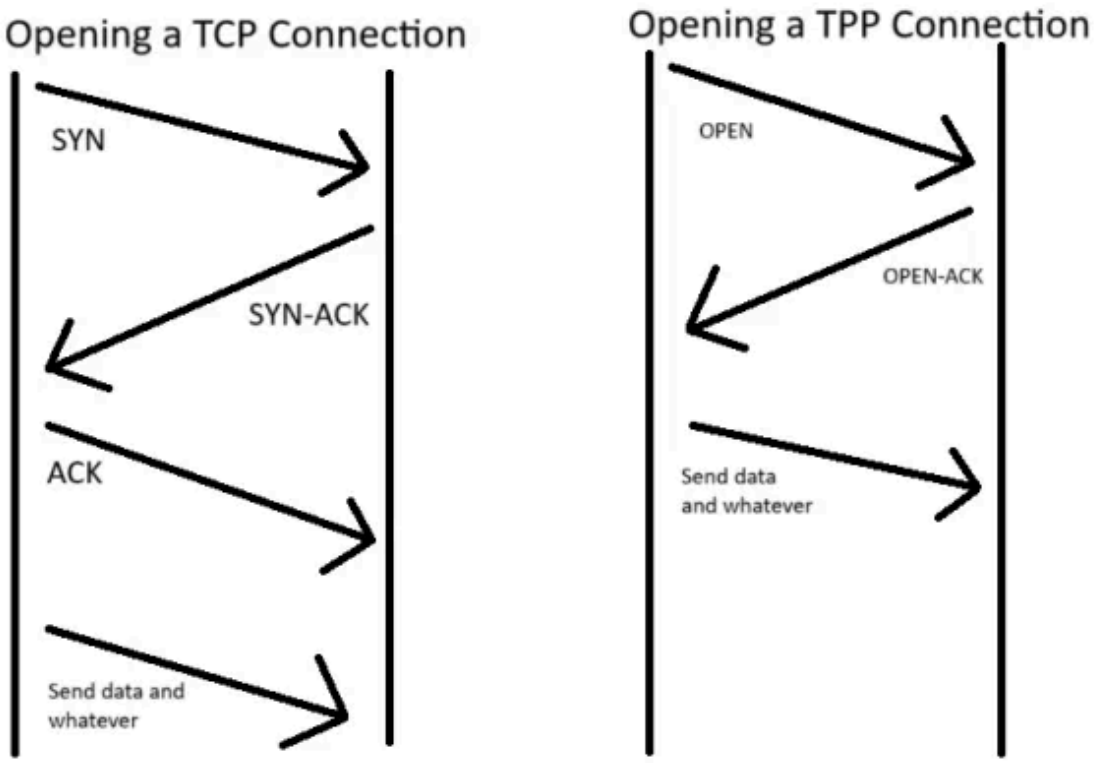
Indeed, every organisation inherits its behaviours from its leader, and the documentation for Tesla's new open-source transport protocol designed to ru over an ethernet network, named the Tesla transport protocol over Ethernet (TTPoE), reads like one of its CEOs tweets/posts on X.



Source: ServeTheHome

At HotChips 2024, Tesla displayed a liquid cooled Dojo training tile, connecte NICs and a network, all on a desk at the conference. Whilst this was an impressive and easy to understand display, what was arguably even more impressive was their announcement of their own high-performing open-sou transport protocol designed to run through standard ethernet switches, or th TTPoE.

Tesla would know just as well as Google, Microsoft and Meta about how diffi and slow it can get to scale ethernet networks to 10s or 100's of thousands of endpoints, but unlike the other tech giants, Tesla has decided to open source transport layer protocol and has designed it to run only over Ethernet, allowi users to avoid InfiniBand whilst seeing the latency reductions and bandwidt increases provided by this new protocol.



Source: Tesla

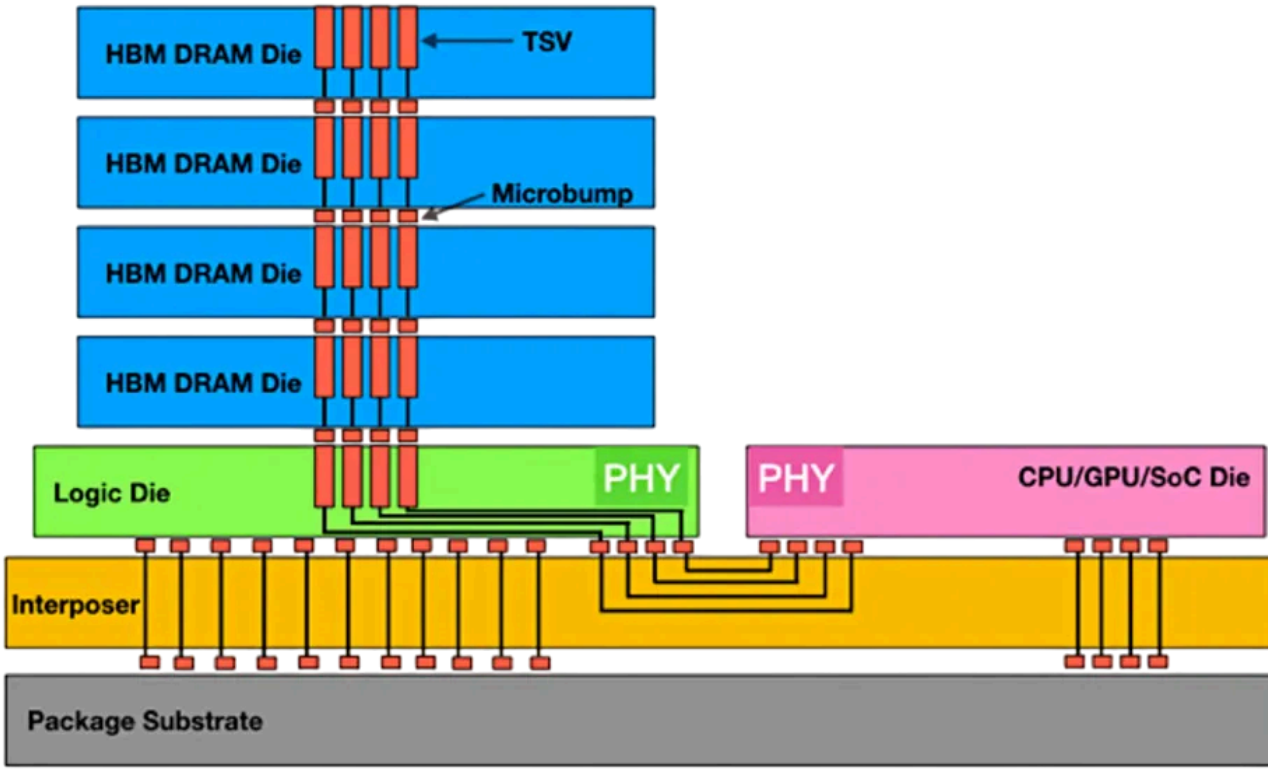
TTPoE improves on TCP for AI and HPC workloads by removing excessive acknowledgements like those at connection closings, removing unnecessary waits, and changing how lost packets and re-transmission is handled in hardware, among many other new features. However, it keeps TCP's endpoin managed congestion control and doesn't require "smart" NICs, emphasising c differences to InfiniBand.

The race towards HBM4

High bandwidth memory, or HBM is the solution that industry has come up for keeping up with the incredible increase in compute power from the highe end GPUs, but this has led to a close race between two competitors. What is H why do we need it, and who will win this race? We answer two of these questions.

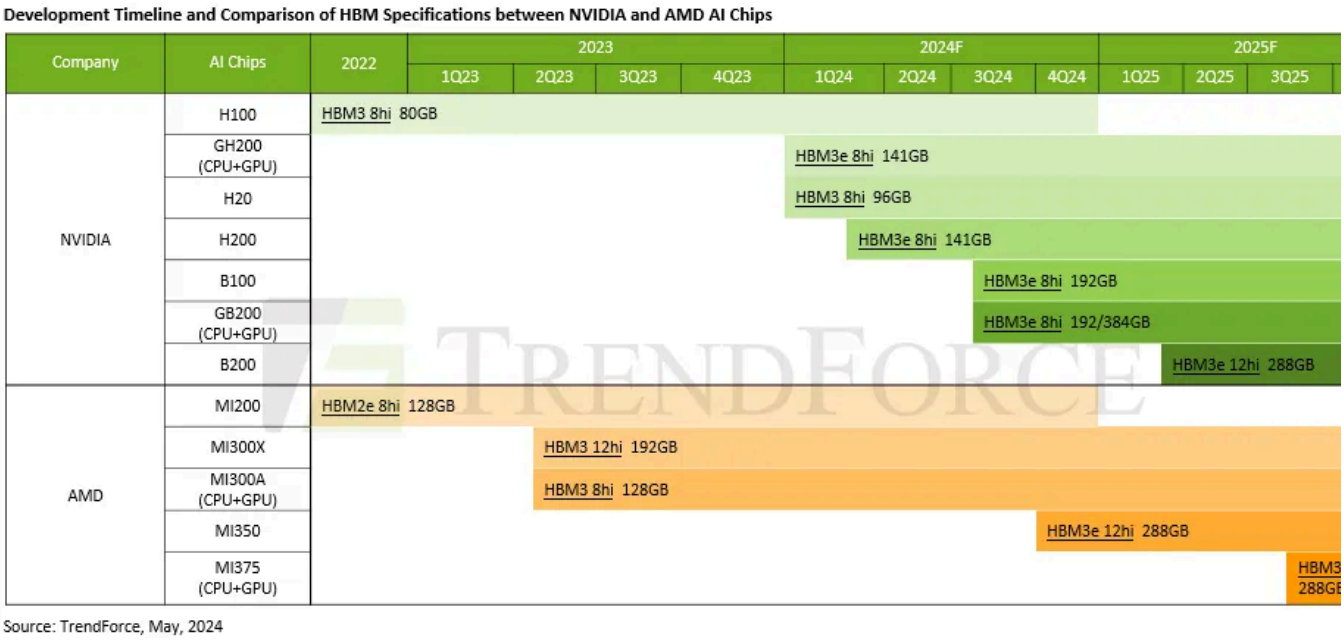
With regular DRAM, you have one die (or block of silicon) with the transistor and capacitors for storing memory, but since compute has been scaling faste

than memory for decades, we have had to stack multiple (specialised) DRAM dies vertically (diagram below) and link them up to access all the dies at onc providing more bandwidth per chip area from memory modules. Currently, tl latest Nvidia H1/200s and AMD MI300X/A ship with HBM3e and HBM3 respectively, third generation HBM modules. Towards the end of 2024 and 2025/6, we will see 12 high stacks for HBM3e and possibly 16 or more for HBM reaching ever higher volumes and bandwidths.



Source: Rambus

It's clear that at least until 2025/6, this will be a two-horse race between Samsung and SK Hynix, both Korean memory manufacturers, with Micron picking up both market share and pace from 2H2024 onwards.



Source: TrendForce

One-pagers:

LPDDR5x

With IBM, Google, Amazon and many others pushing for gains in efficiency rather than raw performance, energy efficient AI hardware is becoming the popular choice for many datacentre users. Low power double data rate (LPDD memory is a key component this drive towards effciency.

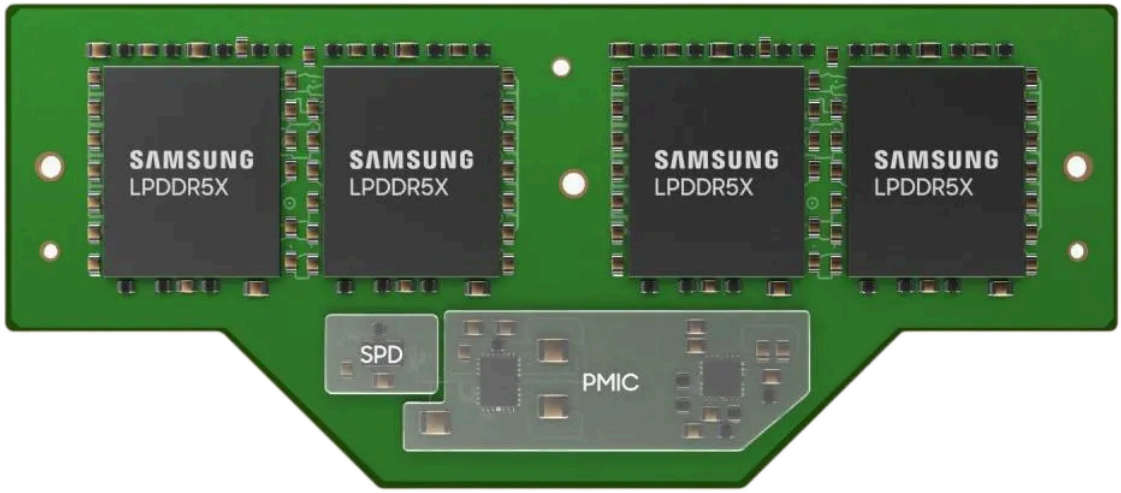
Whilst competitors in the large AI model training space like Nvidia and AMD head towards higher and higher power usage, others are focusing on the AI inference market, where budgets are defined by opex and service utilisation. Here, competing with the power-hungry memory used in training accelerato key, and one of the key components in this drive is LPDDR, generation 5x.

Let's break down what this means:

- LP: Low Power – This is achieved by a lower effective operating power du variability in the voltage required to refresh cell states, as well as smarte

power usage in idle states and finer grain control over which parts of the memory are consuming power at a given time

- DDR: Double Data Rate – A standard feature of DRAM (Dynamic random-access memory, or just main memory), where the rate at which data can transferred is doubled by using by using both the rise and fall of each clo cycle (or tick, like in a clock) to transfer data
- 5x: Generation 5, Enhanced ("x" sounds better than "e" I suppose)



Source: Samsung

LPDDR must be soldered directly onto the mainboard, resulting in an inabilit hot-swap (replace the physical hardware during operation) modules or chan the volume, but for mobile devices and AI accelerators, this has been the case even with regular DRAM. The benefits far outweigh the inconveniences and additional costs through, with lower power usage, lower latency, and higher bandwidth.

OSFP

Despite enabling the generative AI explosion, networking doesn't get the sam media treatment that compute does, with GPUs taking the spotlight. Connect multiple GPU nodes requires fast networks, and hence fast transceivers. Octa Small Form factor Pluggables (OSFP) are the latest in this trend.

Connecting multiple servers together requires a network, and hence network cables. Whether its copper (not widely used anymore) or optical, there needs be a component that can connect to a cable and then plug into a switch/adap whilst mediating physical signal quality and modulation, and also the conversion between electrical/optical signals in wire/fibre and electrical sign understandable by a switch or a server.

SFPs have been in use for a while now, but more recently the need for extrem high bandwidth interconnect between servers has exploded due to the popularisation of training and inferencing very large generative AI models. P to that, users who ran very large workloads like weather forecasting or struct stress analysis could balance their compute and memory requirements with 25Gb or even 10Gb network bandwidths, but with AI, the memory demand is much greater that the technology needed to transfer data between servers ha had to evolve incredibly fast to keep up. In addition, with GPUs like A100s and H100s, the compute speeds have increased rapidly, meaning that memory sp are now more of a bottleneck than ever.

Source: Ascent Optics

A 400G (4x100G) 8-lane optical transceiver (transmitter and receiver pair per lane) unit designed to be plugged into multi-port network switches, hence requiring a small form factor. Within this body, OSFPs have to fit their connectors, signal converters, heat sinks and more. The distances supported optical connections using OSFPs can be up to 10KM, allowing for very long-ra optical connections between datacentres or multiple facilities.

FLOPs

How do you compare different compute hardware? There are a few key metri depending on the purpose of the device, but for CPUs and GPUs, a key metric the number of floating-point operations per second, or how fast you can add multiply lots of numbers. But this can be misleading.

Hardware		FLOPs (Trillions)
Nvidia H100 SXM		1979
Google TPUv4		275
AMD EPYC 7702		2.48*
Cerebras WSE-3		125,000

*Estimated

Looking at these numbers without any context or purpose is meaningless an says very little about any of the devices. For example:

- Nvidia's H100 reaches ~2000 Tera FLOPs, but comes with some small pri only with sparsity, or where most of the data is effectively zero
- Google's TPUv5e looks a lot worse off, but they advertise for dense matric (most of the data is not zero) and the power draw of a TPUv4 is less than quarter that of the H100
- AMD's EPYC 7702 CPU doesn't even advertise FLOPs so it had to be estim since what CPUs do is much more complicated and flexible than what GI do

- Cerebras' WSE-3 looks amazing, apart from when you learn its many, ma times the size and cost of a H100 GPU and is much harder to program/optimise for

Previous Next

Discussion about this post

Comments Restacks

Write a comment...

ThunderMikey 14 Feb

Liked by Hitesh Kumar

Cool content

LIKED (1) REPLY

1 reply by Hitesh Kumar

1 more comment...